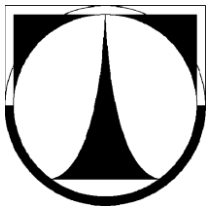
TECHNICKÁ UNIVERZITA V LIBERCI

Ekonomická fakulta



BAKALÁŘSKÁ PRÁCE

2013 **Jitka Ládrová**

TECHNICKÁ UNIVERZITA V LIBERCI
Ekonomická fakulta

Studijní program: B 6209 Systémové inženýrství a informatika
Studijní obor: Manažerská informatika

Praktické využití pokročilých analýz dat pro manažerské rozhodování ve zdravotní pojišťovně

Application of Advanced Data Analysis Methods for Managerial Decisions in the Health Insurance

BP-EF-KIN-2013-14

Jitka Ládrová

Vedoucí práce: Ing. Vladimíra Zádová, Ph.D., KIN
Konzultant: Mgr. Ing. Martin Šály, KOMIX s.r.o.
Počet stran: 55
Počet příloh: 0

Datum odevzdání: 10. 05. 2013

Zadání bakalářské práce

- Bude vloženo

Zadání bakalářské práce

- Bude vloženo

Prohlášení

Byla jsem seznámena s tím, že na mou bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb. o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci (TUL) nezasahuje do mých autorských práv užitím mé bakalářské práce pro vnitřní potřebu TUL.

Užiji-li bakalářskou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědoma povinností informovat o této skutečnosti TUL; v tomto případě má TUL právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Bakalářskou práci jsem vypracovala samostatně s použitím uvedené literatury a na základě konzultací s vedoucím bakalářské práce a konzultantem.

V Liberci, 07. 05. 2013

Poděkování

Je mou milou povinností poděkovat všem, s jejichž pomocí tato bakalářská práce vznikla.

Děkuji paní Ing. Vladimíře Zádové Ph.D. za odbornou pomoc a vedení. Další velký dík patří i mému konzultantovi, kterým byl pan Mgr. Ing. Martin Šály, za jeho trpělivost a ochotu.

Rovněž bych na tomto místě ráda poděkovala rodičům, blízkým a firemním kolegům, že mě vždy během tvorby bakalářské práce podrželi a stáli po celou dobu při mně.

Anotace

Bakalářská práce, jak již její název napovídá, se zabývá praktickým využitím pokročilých analýz dat, respektive data miningu coby disciplíny, do které se řadí většina metod, které se pro zmiňované analýzy využívají. V bakalářské práci je tudíž řešen teoretický rámec data miningových metod, na které navazuje jejich použití pro projekt zdravotní pojišťovny a následná aplikace výsledků pro manažerské rozhodování. Projekt zdravotní pojišťovny, který je popisován v praktické části, měl pak za cíl odhalit podvodně vykazovanou zdravotní péči některých zdravotnických zařízení, respektive vytipovat mezi všemi lékaři vedenými v databázi zdravotní pojišťovny ty, kteří vykazují prováděné výkony nestandardním způsobem a u kterých je tudíž vysoká pravděpodobnost toho, že se neoprávněně obohacují na úkor zdravotní pojišťovny.

Klíčová slova

Data mining, využití data miningu, analýza dat, zdravotní pojišťovna, manažerské rozhodování

Annotation

The bachelor thesis, as its name suggests, deals with practical application of advanced data analysis, or data mining as a discipline, to which belong most of the methods that are used for the before-mentioned analysis. There is a theoretical framework for data mining methods designed in the bachelor thesis, followed by their use for a health insurance project and subsequent applications for managerial decision-making. The health insurance project, which is described in the application part, should aim to detect fraudulent reported activities of some health care facilities, or to identify among all doctors listed in the database of health insurance those that report performance of acts abnormally and for which is therefore a high probability of unduly enriching at the expense of health insurance.

Keywords

Data mining, the use of data mining, data analysis, health insurance, managerial decision making

OBSAH

ÚVOD **11**

ZHODNOCENÍ SOUČASNÉHO STAVU **12**

1 TEORETICKÁ ČÁST **14**

1.1 DATA MINING 14
1.2 HISTORIE DATA MININGU 15
1.3 DATA MINING A JEHO VLIV 16
1.4 TYPY DATA MININGOVÝCH ÚLOH 18
1.4.1 Úlohy *21*
1.4.2 Tři zdroje *23*
1.4.3 Metody *26*
1.5 NÁSTROJE PRO DATA MINING 29
1.5.1 Charakteristiky vybraných nástrojů *29*
1.5.2 Výběr nástroje *32*

2 PRAKTICKÁ ČÁST **33**

2.1 VYUŽITÍ DATA MININGOVÝCH ÚLOH PRO MANAŽERSKÉ ROZHODOVÁNÍ VE ZDRAVOTNÍ POJIŠŤOVNĚ 33
2.1.1 Nejpoužívanější data miningové metody *33*
2.1.2 Metody použité pro vybranou zdravotní pojišťovnu *34*
2.2 PŘEDMĚT PLNĚNÍ ZAKÁZKY 35
2.3 ZÍSKÁNÍ A PŘÍPRAVA DAT 36
2.3.1 Získání dat *36*
2.3.2 Příprava dat *37*
2.4 ANALÝZA DAT 39
2.4.1 Analyzovaná data *40*
2.4.2 Modely *40*
2.5 VÝSLEDKY DATOVÉ ANALÝZY A ZHODNOCENÍ PŘÍNOSŮ BAKALÁŘSKÉ PRÁCE 48
2.5.1 Výsledky datové analýzy *48*
2.5.2 Zhodnocení přínosů *51*

ZÁVĚR **52**

SEZNAM POUŽITÉ LITERATURY **54**

SEZNAM OBRÁZKŮ

Obrázek 1 Klasifikace, predikce 21
Obrázek 2 Deskripce 22
Obrázek 3 Hledání nuggetů 22
Obrázek 4 Rozhraní nástroje SPSS Modeler 30
Obrázek 5 Rozhraní nástroje SAS 31
Obrázek 6 Počty hodnot IPD 50

SEZNAM TABULEK

Tabulka 1 Typy úloh 1 19
Tabulka 2 Typy úloh 2 19
Tabulka 3 Kontingenční tabulka 24

SEZNAM POUŽITÝCH ZKRATEK, ZNAČEK A SYMBOLŮ

EIS ..... Executive Information Systems

IPD ..... index potenciálního darebáctví

OLAP ..... Online Analytical Processing

QBE ..... Query By Examples

SQL ..... Structured Query Language

Úvod

Hlavním cílem a přidanou hodnotou, kterou by bakalářská práce měla přinést, je praktické využití analýz dat, a to konkrétně těch, které spadají do oblasti data miningu. Ty totiž v drtivé většině bývají popsány pouze v teoretické rovině, nebo se o jejich použití mluví velice obecně.

Bakalářská práce je zaměřena na jednu ze specifických oblastí, ve které se pohybuje velké množství dat. Touto oblastí je zdravotnická péče, respektive v konkrétním případu, který je rozebírán v praktické části, analýza dat jedné ze zdravotních pojišťoven působících na českém trhu.

Jak je všeobecně dobře známo, kolem zdravotnictví a zdravotních pojišťoven se pohybují nemalé částky peněz, což logicky vede k tomu, že se najdou tací, kteří této skutečnosti chtějí využít ve svůj prospěch. Napomáhá jim v tom právě fakt, že zdravotní pojišťovny pracují s velkými objemy dat a pomocí běžných kontrol je prakticky nemožné drobné, ale často opakované prohřešky vysledovat, protože se „ztratí“ mezi nemalým množstvím jiných úhrad.

Zde se proto nabízí otázka, jak je možné podobnému chování zabránit. Odpověď lze nalézt právě v metodách dolování dat. Dalším cílem bakalářské práce je tudíž poskytnout řešení tohoto problému a popsat, jak se konkrétně na daném projektu zdravotní pojišťovny abnormality, které indikují podvodné vykazování lékařských úkonů, zjišťovaly.

Bakalářská práce je rozdělena do dvou základních celků, teoretické a praktické části.

V teoretické části je nejprve vysvětlen pojem data mining, s nímž jsou pokročilé analýzy dat spjaty. Dále pak jsou v této části práce popsány jednotlivé metody a nástroje, které se pro dolování dat používají.

Do druhé, praktické části práce pak spadá popis projektu pro zdravotní pojišťovnu, příprava dat, výběr metod a nástrojů popsaných v teorii pro tento konkrétní případ a samozřejmě popis jednotlivých modelů, které byly použity. Praktickou část uzavírá vyhodnocení výstupů a jejich následné využití pro manažerské rozhodování.

Zhodnocení současného stavu

Cílem zhodnocení současného stavu je podat informace o tom, jak je problematika spojená s tématem bakalářské práce řešena v již známých případech a jak je o ní psáno v doposud publikovaných zdrojích.

Hned na samém začátku je nutné říct, že pokud se v dnešní době mluví o pokročilých analýzách dat, myslí se tím v drtivé většině případů metody, které jsou součástí *dolování dat*, tzv. data miningu. A právě na data mining se jak nyní v literární rešerši, tak i následně v bakalářské práci samotné, zaměříme.

O data miningu a jeho úlohách a metodách pojednává nemálo odborných publikací. Pokud bychom měli vybrat a zmínit alespoň některé z nich, byly by to knihy *Data Mining: Concepts and Techniques* od dvojice autorů Jiawei Han a Micheline Kamber; *Data Mining: Practical Machine Learning Tools and Techniques* od Marka Halla, Iana Wittena a Eiba Franka; z česky psaných publikací by to pak byla kniha Petra Berky *Dobývání znalostí z databází*.

Všechny tyto knihy se věnují data miningu sice značně podrobně, ale zároveň do velké míry čistě obecně. Bližší informace o aplikaci data miningových úloh v manažerském rozhodování, natož praktické příklady jejich použití, v nich nenalezneme.

Pokud budeme v hledání literatury pokračovat o něco konkrétněji a zaměříme naši pozornost přímo na využití data miningu v ekonomických oborech, potažmo právě v managementu a manažerském rozhodování, nalezneme publikace, které se dané problematice věnují již hlouběji. Jmenovitě to jsou knihy *Data Mining Techniques For Marketing, Sales, and Customer Relationship Management* od spoluautorů Michaela J. Berryho a Gordona S. Linoffa; a *Data Mining and Statistics for Decision Making* od francouzského odborníka Stéphana Tufféryho.

Když ovšem zamíříme ještě o něco hlouběji, a to přímo k aplikaci data miningu pro potřeby zdravotní pojišťovny, zjistíme, že v dostupných zdrojích žádné publikace ani odborné články zpracovávající toto téma nejsou. Jediné, co můžeme v tomto ohledu nalézt, jsou zahraniční prezentace (jako například [1]) s kvalifikovanými odhady, kolik by šlo data miningovými úlohami v tomto odvětví ušetřit, kde se pohybují jaké částky v milionech

dolarů a podobně. Postupy jak těchto výsledků dosáhnout v nich jsou sice nastíněny, ale pouze velice obecně. Pokud se na to podíváme z manažerského hlediska, je to pochopitelné. Jakákoliv konkrétní aplikace pokročilých analýz je pro každou firmu, která se touto problematikou zabývá, cenné know-how, a to si celkem po právu chrání.

1 TEORETICKÁ ČÁST

V první části práce věnované teoretickému rámci problematiky týkající se praktického využití pokročilých analýz dat se budeme zabývat ryze obecnými poznatky.

Jak již bylo v předešlé části nastíněno, pokud v současné době hovoříme o pokročilých analýzách dat, jsou tím myšleny především úlohy a metody data miningu.

V kapitole *1 Teoretická část* proto bude vysvětlen pojem data mining, okrajově se podíváme do jeho historie, zamyslíme se nad možným dopadem, který pro nás jako lidstvo má, a poté rozebereme dělení data miningu podle jednotlivých úloh a metod.

Jako poslední podkapitola části věnované teorii bude uveden seznam nástrojů, které se v současné době pro získávání informací z datových zdrojů využívají. Součástí dané podkapitoly bude stručný popis těchto nástrojů a jejich charakteristiky, které sloužily pro následný výběr pro praktickou část bakalářské práce.

1.1 Data mining

Data mining má jako většina pojmů spojených s počítači a obecně IT svůj původ v anglickém jazyce. I když se výraz data mining používá zcela běžně i v češtině, jeho český ekvivalent samozřejmě existuje. Nejvíce rozšířeným a používaným překladem slovního spojení *data mining* je označení *dolování dat*, tedy doslovné přeložení pojmu. Dalším možným synonymem je pak *vytěžování (z) dat* či výraz *dobývání znalostí*.

Ať už data mining budeme označovat jakkoliv, jeho podstatu to nijak neovlivní. Stále půjde o souhrn analyticko-metodologických postupů, díky kterým je možné dobrat se i k netriviálním informacím, které bývají v datech velice často na první pohled skryté.

Poslední zmiňovaná informace nás přenáší k možným definicím pojmu data mining. A jak už množné číslo slova definice poukazuje, existuje více různých způsobů, jak dolování dat obecně definovat.

První, kratší a jednodušší, definicí dle [2 s. 149] je:

„Data mining je hledání hodnotných informací ve velkých objemech dat.“

Pokud bychom pak pátrali po více specifické definici, dle stejného zdroje [2 s. 149] bychom nalezli následující:

„Data mining je netriviální proces zjišťování platných, neznámých, potenciálně užitečných a snadno pochopitelných závislostí v datech.“

Jak už bylo v první definici nadneseno, v rámci data miningu se pracuje se soubory dat o velkých objemech. Není vůbec žádnou výjimkou, že databáze, nad nimiž se následně jednotlivé níže rozebírané metody (podkapitola *1.4.3 Metody*) provádí, jsou v řádech desítek miliónů řádků. To byl ostatně i jeden z hlavních důvodů, který dal data miningu vůbec vzniknout, neboť pracovat se stále většími a obsáhlejšími databázemi nebylo už bez jeho pomoci možné.

Hlavní hybnou silou, proč vůbec bylo zapotřebí nějakým způsobem analyzovat data, která v každé firmě vznikají a které si společnosti v určité podobě uchovávají, byly beze sporu peníze. Ty se v dnešním světě skrývají takřka za vším a u vzniku data miningu tomu nebylo jinak. Primárním cílem dolování dat je zajistit konkurenceschopnost podniku a zvýšit tak jeho zisk.

1.2 Historie data miningu

Počátky toho, co dnes nazýváme data miningem, se datují zhruba od 60. let 20. století. Jako první se pro tehdejší dobývání znalostí začaly využívat regresní analýzy (více v podkapitole *1.4.2 Tři zdroje – Statistika – Regresní analýza*), které následovaly rozhodovací stromy (podkapitola *1.4.3 Metody – Rozhodovací stromy*).

Dlouho dobu byl zájem o data mining víceméně pouze v akademické rovině, neboť mnozí lidé pochybovali o jeho praktickém využití. Průlom nastal až v devadesátých letech. Tehdy se data některých organizací přenesla přes únosnou hranici, kdy už se k informacím potřebným pro manažerské rozhodování nebylo možné dostat pomocí v té době běžných tabelačních metod. V té samé době byly objeveny rovněž způsoby, jak zamezit tzv.

falešným korelacím[1], tedy případům, kdy dle [3] dvě složky *m*-rozměrného náhodného vektoru silně korelují (tj. vzájemně souvisí) s neuvažovanou třetí náhodnou veličinou.

V současné době je data mining již uznávanou metodou, jak se k těžko přístupným informacím z dat dostat. I tak ale zatím není natolik rozšířený, jak by se možná mohlo na první pohled zdát. Alespoň tedy v našem českém měřítku tomu tak zdaleka není.

Zakázek pro dolování dat je na domácím trhu v tomto okamžiku pouze poskrovnu. Jednou z mála výjimek jsou velké nadnárodní korporace, které data mining využívají a o nichž se bude více hovořit v následující kapitole.

1.3 Data mining a jeho vliv

Jak již bylo nadneseno, data mining vznikl především jako cílené řešení problému, jak z rozsáhlých databází získat užitečné informace pro manažerské rozhodování, a tím pádem jak uspořit nemalé finanční prostředky. Pokud budeme z tohoto vycházet, může nám být hned jasné, že jeho vliv je nemalý.

Data mining zasahuje do různých oblastí. Jako nejčastější příklady jeho využití se uvádí v prvé řadě přímý marketing, kde se pomocí dostupných dat analyzuje, které klienty dané společnosti by bylo vhodné vybrat pro oslovení například s novým produktem či službou, aby pravděpodobnost úspěchu nabídky byla co nejvyšší.

Jako další segment, kam data mining nemalou měrou zasahuje, bychom mohli jmenovat prodej v maloobchodech, především v sítích velkých prodejních řetězců. Ty pro zvýšení svých výnosů používají data mining pravidelně, protože objemy dat, se kterými se v tomto odvětví pracuje, jsou jedny z největších a také nejčastěji měnících se. Pro demonstraci možných výkyvů můžeme zmínit například stále častější potravinové aféry, které díky medializaci nabírají velkých rozměrů a nemalou měrou tak výsledky datových analýz ovlivňují.

[1] Příkladem, kdy může jev zvaný falešné korelace nastat, je databáze, ve které se vyskytují chybějící hodnoty.

Pokud ovšem nyní pomineme mediální kauzy a zaměříme se čistě na chování řetězců v oblasti sbírání dat o svých zákaznících, zjistíme, že drtivá většina velkých obchodů je v tomto ohledu velice chytrá. Jejich manažeři si jsou dobře vědomi, že nejcennějšími prostředky jak maximalizovat své zisky není nic jiného než právě informace o jejich zákaznících. Proto se stále častěji setkáváme s nejrůznějšími obdobami věrnostních klubů. Člověk poskytne o sobě obchodu základní informace, získá kartu dané společnosti či řetězce, kterou pak při každém nákupu předloží, aby získal určitou slibovanou výhodu, kvůli které si členství zřizoval. Taková je cena, za kterou si obchody kupují informace o svých zákaznících. Díky tomu jsou pak jejich manažeři schopni velice přesně určit, co daný člověk nejčastěji nakupuje, jak často, kde apod., a i když se to může zdát jako ne moc užitečné informace, obchody takto mohou lépe cílit nabídky, nebo například vysledovat v kolik hodin a v jaké dny bývají nejvíce/nejméně vytížené pokladny a podle toho přemístit potřebný personál na místo, kde bude pracovní síla nejlépe využita.

Vzhledem k právě zmíněným faktům je jasné, že data mining v sobě nese skrytou hrozbu v podobě jakési nesvobody a možného zneužití citlivých osobních údajů. Mnozí lidé si ani neuvědomují, co všechno na sebe prozrazují, aniž by o tom vlastně věděli. Pod povrchem toho všeho je schované jisté vyšší morální hledisko, zda je něco takového opravdu správné, vhodné, etické – zkrátka, je-li to skutečně ta nejlepší cesta, kterou bychom se měli a chceme ubírat.

Obecně se na tuto otázku odpovědět nedá, protože co člověk, to jiný názor. Na druhou stranu každý z nás velice dobře ví, že nic není zadarmo, a pokud doopravdy stojíme o to, abychom něco dostali levněji, nebo třeba za odměnu, musíme si předem uvědomit, že to svou jistou cenu mít bude, i když v tomto případě ne peněžní. Navíc rozhodně se nedá říct, že by data mining představoval jen a pouze negativa, jak to nyní může vypadat.

Velice často se můžeme setkat s tím, že dolování dat slouží jako prostředek k prevenci. Byly zaznamenány případy, kdy se za využití pokročilých datových analýz v databázích letišť a leteckých společností našla jména teroristů připravujících se na útok. Na podobnou věc by se za normálních okolností nikdy nepřišlo, neboť zmiňovaná skupina byla odhalena díky tomu, že analytici v datech objevili velice nepravděpodobnou „náhodu“, a sice že jistí tři muži, kteří letěli z různých míst a opět do odlišných zemí, přestupovali už po třetí za

sebou všichni ve stejné dny na stejných letištích. A to určitě není jediný příklad toho, kdy data mining slouží pro dobré účely.

Dalším odvětvím, kde se dá data mining využívat v tom lepším slova smyslu, je již v názvu bakalářské práce zmiňovaný sektor zdravotní péče a pojišťoven. Zde, jak bude popsáno v praktické části, může dolování dat pomoci (a v praxi doopravdy i pomáhá) při odhalování podvodně vykázaných lékařských výkonů.

1.4 Typy data miningových úloh

Úlohy z oblasti data miningu se dají dělit vícero způsoby a není jednoznačně řečeno, jaký pohled je ten nejlepší, nejsprávnější či nejuniverzálnější.

Vysvětlení, proč tomu je právě takto, by se dalo najít jistě hodně. Tím nejpravděpodobnějším je fakt, že data mining je poměrně mladá a rychle se rozvíjející oblast informatiky, a tudíž ještě neuběhl dostatek času na to, aby některé dělení získalo výrazně navrch před ostatními.

První dělení dle [4] uvedené v *tabulce 1 Typy úloh 1* navrhl S. M. Weiss. Toto rozdělení řadí úlohy data miningu do dvou kategorií, a to do úloh kategorie *predikce* (čili úlohy, které nějakým způsobem ukazují to, jaké hodnoty jsou očekávány podle dat předchozího vývoje) a do úloh *deskriptivní analýzy* (tedy úloh, při kterých dochází k objevování znalostí a poznávání nových faktů).

Predikce	Objevování znalostí (deskriptivní analýzy)
Klasifikace Regresní analýza Analýza časových řad	Zjišťování odchylek Segmentace databáze Shlukování Asociační pravidla Sumarizace Vizualizace Dolování v textu

Tabulka 1 Typy úloh 1
Zdroj: http://datamining.xf.cz/view.php?cisloclanku=2002102801

Druhé dělení opět dle stejného zdroje [4] (viz *Tabulka 2 Typy úloh 2*) v sobě zahrnuje kromě samotných úloh i jednotlivé druhy metod, pomocí nichž se dané konkrétní úlohy řeší.

Úloha	Metoda
Klasifikace	Diskriminační analýza
	Logistická regresní analýza
	Rozhodovací stromy
	Neuronové sítě (back propagation)
Odhady hodnot vysvětlované proměnné	Lineární regresní analýza
	Nelineární regresní analýza
	Neuronové sítě (radial basis function)
Segmentace	Shluková analýza
	Genetické algoritmy
	Neuronové shlukování (Kohonenovy mapy)
Analýza vztahů	Asociační algoritmus pro odvozování pravidel typu If X, then Y
Predikce v časových řadách	Boxova-Jenkinsova metodologie
	Neuronové sítě (recurrent back propagation)
Detekce odchylek	Vizualizace
	Statistické postupy

Tabulka 2 Typy úloh 2
Zdroj: http://datamining.xf.cz/view.php?cisloclanku=2002102801

Obě dvě dělení poskytují zajímavý a trochu odlišný pohled na to, jak lze na data mining jako takový nahlížet.

Pokud se pak na oba dva typy dělení podíváme blíže, objevíme ukázku toho, jak hranice mezi úlohou a metodou může být velice tenká. Toto tvrzení dokládá konkrétně vizualizace, u které si lze povšimnout, že v prvním Weissově rozdělení je řazena přímo mezi úlohy, kdežto v t*abulce 2 Typy úloh 2* ji nalezneme mezi jednotlivými metodami pro detekci odchylek.

Posledním dělením, které bude v bakalářské práci rozebráno více dopodrobna a ze kterého budou následně vybrány konkrétní metody pro manažerské rozhodování, vychází z [5].

V této publikaci je nahlíženo na dolování dat poněkud netradičním způsobem, a to podle toho, jak se data mining postupně vytvářel a jednotlivé kroky zdokonalovaly.

Jako první se autor zmiňuje o úlohách jako takových, kteréžto dělí na tři skupiny:

- klasifikace nebo predikce
- deskripce
- hledání „nuggetů"

Na toto dělení navazují tzv. *tři zdroje*, do kterých se řadí:

- databáze
- statistika
- strojové učení

Posledním stupněm v této pomyslné pyramidě pak jsou metody modelování:

- rozhodovací stromy
- asociační pravidla
- rozhodovací pravidla
- neuronové sítě
- evoluční algoritmy
- Bayesovská klasifikace
- metody založené na analogii
- induktivní logické programování

1.4.1 Úlohy

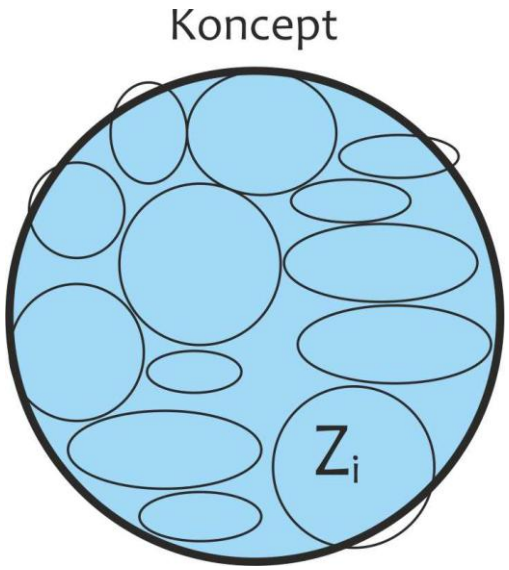
V této podkapitole budou nyní více přiblíženy jednotlivé typy úloh z třetího, posledně zmiňovaného, způsobu dělení.

Klasifikace, predikce

Hlavním rozdílem mezi klasifikací a predikcí je časové hledisko, které právě u zmíněné predikce hraje klíčovou roli. Zatímco klasifikaci bychom mohli označit jako ohodnocení daných dat, predikce znamená předpověď nebo výhled. Z těchto pojmů intuitivně cítíme, že se jedná o něco, co se týká budoucnosti.

Cíl klasifikace, respektive predikce, je nalezení znalostí použitelných pro hodnocení nových případů. Důraz je pak kladen především na to, aby získané znalosti odpovídaly předem zadanému konceptu.

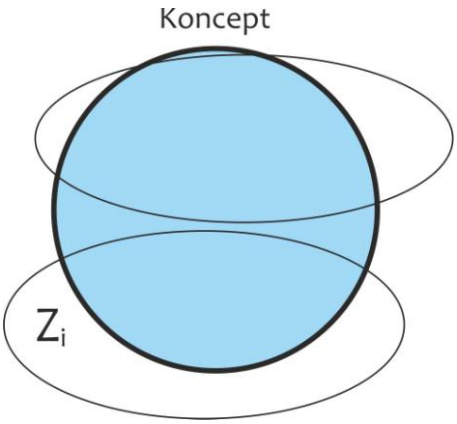
Základním principem tohoto typu úlohy je, že přesnost pokrytí má přednost před jednoduchostí, čili dle [5 s. 18] *„připouštíme větší množství méně srozumitelných dílčích znalostí“*, jak můžeme vidět na *obrázku 1 Klasifikace, predikce*.



Obrázek 1 Klasifikace, predikce
Zdroj: BERKA, P. Dobývání znalostí z databází. 1. vyd. Praha: Academia, 2003. ISBN 80-200-1062-9.

Deskripce

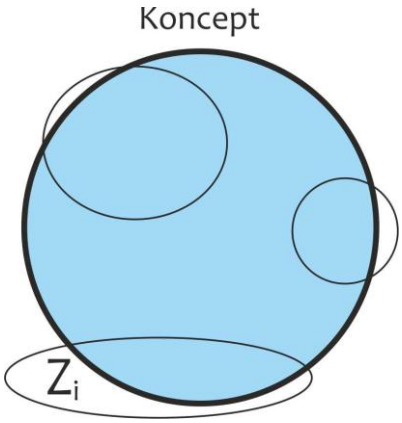
Pojem deskripce bychom mohli vysvětlit jako popis, jehož cílem je objevit dominantní struktury nebo vazby, jež jsou v daných datech skryté a na první pohled neviditelné.
V úlohách typu deskripce je hlavním požadavkem srozumitelnost znalostí pokrývajících daný koncept. Při deskripci tedy, když budeme znovu citovat [5 s. 18] *„dáváme přednost menšímu množství méně přesných znalostí“*. Grafické zobrazení deskripce pak můžeme vidět na *obrázku 2 Deskripce*.



Obrázek 2 Deskripce
Zdroj: BERKA, P. Dobývání znalostí z databází. 1. vyd. Praha: Academia, 2003. ISBN 80-200-1062-9.

Hledání „nuggetů“

Hledání „nuggetů“ znamená nacházení určitých vzorů a pravidel, od nichž požadujeme, aby přinášely zajímavé, překvapivé či v něčem novém znalosti. Hlavní rozdíl od předchozích úloh je, že od takto nalezených poznatků není požadováno, aby plně pokrývaly daný koncept, jak naznačuje *obrázek 3 Hledání nuggetů*.



Obrázek 3 Hledání nuggetů
Zdroj: BERKA, P. Dobývání znalostí z databází. 1. vyd. Praha: Academia, 2003. ISBN 80-200-1062-9.

1.4.2 Tři zdroje

Jak již bylo výše uvedeno, v této podkapitole se budeme zabývat třemi zdroji, na něž pak dále navazují jednotlivé metody.

Databáze

Po prvních databázích, ve kterých dle [5] byla data ukládána do jednoho velkého souboru, se objevily tzv. relační databáze, které byly velice významné pro následný vývoj data miningu.

Relační databáze je tvořena množinou relací (tabulek), které jsou mezi sebou navzájem propojené. Informace z takto uložených dat lze získat pomocí dotazovacích jazyků QBE (Query by Example) nebo SQL (Structured Query Language), přičemž jazyk SQL je standard pro definici a manipulaci dat uložených v relačních databázích.

Na relační databáze navázal systém EIS (Executive Information Systems), který by se dal zároveň považovat za první krok ke spojení nových technologií s manažerskými aktivitami. Díky uživatelsky přívětivému rozhraní se systém hodil nejen pro programátory, ale také pro manažery, kteří neznali strukturu databáze.

Hlavní nevýhodou EIS byla nedostatečná flexibilita, neboť uživatel-manažer měl k dispozici pouze jistou sadu předpřipravených dotazů. V momentě, kdy se uživatel rozhodl zeptat na něco, co programátor EIS nepředpokládal, musel jít přímo za tvůrcem programu, který následně požadovaný „nestandardní“ dotaz do jazyka SQL převedl.

Problém vyřešila až technologie OLAP, která uživatelům nabídla jak potřebnou flexibilitu, tak příjemné a intuitivní ovládání. Velikým přínosem technologie OLAP je rovněž možnost vizualizace, kdy si uživatel může přehledně zobrazit data nejen v číselné podobě, ale i v té grafické.

Posledním velkým krokem od databází k dobývání znalostí jako takovému (tedy k fázi, kdy nevíme, co přesně chceme hledat) jsou systémy, které umožňují pokládat dotazy na tzv. *asociační pravidla*. Tyto systémy používají algoritmus pro hledání asociací, čili hledají různé podobností mezi jednotlivými daty.

Statistika

Statistika jako věda poskytuje řadu ověřených metod pro data mining. Mezi ty nejvýznamnější a nejčastěji používané se řadí:

- kontingenční tabulky
- regresní analýza
- diskriminační analýza
- shluková analýza

Kontingenční tabulky

Kontingenční tabulky se používají pro přehledné zobrazení vztahu mezi dvěma veličinami, z nichž alespoň jedna je slovní. V kontingenční tabulce pak platí, že řádky odpovídají hodnotám prvního znaku a sloupce pak hodnotám znaku druhého, jak můžeme vidět v následující tabulce z [6], kde:

- n sdružené (simultánní) absolutní četnosti
- $_{i\bullet}n$, $_{j}n_{\bullet}$ okrajové (marginální) absolutní četnosti
- $_{ij}p$ sdružené relativní četnosti
- $_{i\bullet}p$, $_{j}p_{\bullet}$ marginální relativní četnosti.

y_j / x_i	y_1	y_2	. . .	y_s	Součty četností $n_{i\bullet}$
x_1	n_{11}	n_{12}	. . .	n_{1s}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	. . .	n_{2s}	$n_{2\bullet}$
. . .	. . .	. . .	. . .	. . .	. . .
x_r	n_{r1}	n_{r2}	. . .	n_{rs}	$n_{r\bullet}$
Součty četností $n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	. . .	$n_{\bullet s}$	n

Tabulka 3 Kontingenční tabulka
Zdroj: http://multiedu.tul.cz/~katerina.gurinova/multiedu/Statistika_II/Analyza_zavislosti.pdf

Regresní analýza

Regresní analýza navazuje na oblast statistiky zvanou korelace. Korelační analýza se ale na rozdíl od regrese při dolování dat nepoužívá, protože řeší pouze, zdali jsou dvě dané numerické veličiny na sobě lineárně závislé, či nikoliv.

Regresní analýza oproti tomu vychází z předpokladu, že závislost existuje, a zabývá se hledáním konkrétních parametrů pro rovnici této závislosti. Obecný tvar rovnice regresní analýzy vypadá následovně:

(1) $y = q_1x + q_0 + \varepsilon$,

kde parametry q_0 a q_1 jsou hledané veličiny a parametr ε vyjadřuje náhodnou složku, která někdy bývá označovaná i jako tzv. rušivá.

Diskriminační analýza

Diskriminační analýza hledá závislosti jedné nominální veličiny na *x* numerických veličinách. Co se týče statistického členění, diskriminační analýza patří do metod tzv. vícerozměrné (nebo také mnohorozměrné) statistiky a jejím základním cílem je dle [7] rozlišit objekty z konečného počtu tříd na základě objektů z jisté podmnožiny všech objektů.

Shluková analýza

Principem shlukové analýzy je vytvoření skupin objektů, tzv. shluků, které si budou v něčem vzájemně podobné, blízké.

Pokud mluvíme o shlukové analýze, máme tím na mysli celý soubor metod, který se pod tímto označení skrývá [8]. Bližší rozdělení ale není předmětem bakalářské práce, a proto se tímto problémem nebudeme více zabývat.

Strojové učení

Základní myšlenkou strojového učení je předpoklad, že každý organismus, tedy v přeneseném slova smyslu i systém, je schopen se do jisté míry přizpůsobovat měnícím se podmínkám nebo se nějakým způsobem učit z vlastních zkušeností.

Jednotlivé principy, které se v systémech pro získávání znalostí používají, byly převzaty z různých disciplín, kterými dle [5] jsou:

- statistické metody (metody zmíněné v podkapitole Statistika)
- symbolické metody umělé inteligence (rozhodovací stromy, asociační a rozhodovací pravidla)
- subsymbolické metody umělé inteligence (neuronové sítě, genetické algoritmy, bayesovské metody)

1.4.3 Metody

Metody pokročilých analýz dat využívané pro manažerské rozhodování se více méně překrývají se všemi, které se uvádějí v běžném členění data miningu. V této podkapitole bude blíže popsán princip těchto metod.

Rozhodovací stromy

Metoda rozhodovacích stromů je jednou z nejvíce uváděných metod data miningu vůbec. Tento fakt je způsobený tím, že rozhodovací stromy jsou snadno pochopitelné, a proto se dávají velmi často jako ilustrační příklad pro „obyčejné“ uživatele systémů, kteří si normálně pod pojmem dolování dat nedokážou nic konkrétního představit.

Pro vytváření rozhodovacích stromů se používá metoda, při které se data tříští na stále menší a menší podmnožiny, tzv. uzly stromu. Cílem dělení je klasifikace dat, čili nám jde o to, aby po skončení procesu tvorby rozhodovacího stromu převládly ve vzniklých podmnožinách příklady jedné třídy, tj. s jednou konkrétní vlastností.

Asociační pravidla

Princip asociačních pravidel v sobě zahrnuje konstrukci, která je známá ze všech programovacích jazyků – *if-then*.

Na metodě asociačních pravidel je založena jedna z vůbec nejznámějších a v praxi nejpoužívanějších analýz, tzv. analýza nákupního košíku. Ta se zabývá pozorováním chování nakupujících zákazníků a sleduje, co daný spotřebitel koupí, pokud koupí nějakou určitou věc. Například pokud se v zákazníkově košíku objeví pečivo, s vysokou pravděpodobností se v něm vyskytne i máslo.

Cílem této metody je tedy nějakým způsobem předvídat zákazníkův pohyb po prodejně a na základě toho pak třeba upravit rozložení obchodu tak, aby byl maximalizován zisk.

Rozhodovací pravidla

Metoda rozhodovacích pravidel je podobně jako rozhodovací stromy založena na klasifikaci daných dat. Nejde tedy o hledání souvislostí mezi hodnotami atributů a jejich kombinacemi jako u předešlé metody asociačních pravidel, jak by se možná mohlo na první pohled zdát.

Rozhodovací pravidla používají algoritmy *učení s učitelem*, jejichž cílem je naučit se jak správně zařadit jednotlivé příklady do různých tříd.

Neuronové sítě

Princip neuronových sítí jako metody data miningu je velice podobný tomu, jak ve skutečnosti funguje nervová soustava v lidském těle.

Neuronová síť se skládá z umělých neuronů, které jsou mezi sebou navzájem propojené. Díky tomuto spojení si neurony stejně jako v živém organismu předávají signály. Důležitou vlastností každého neuronu pak je, že mohou mít libovolně velký počet vstupů, ale vždy mají jen jeden výstup.

Evoluční algoritmy

Metoda evolučních algoritmů je další z těch, která má svou předlohu v biologických principech. Inspirací pro evoluční algoritmy, jak už název napovídá, byl mechanismus popsaný Darwinem jako přirozený výběr - selekce.

Mezi evoluční algoritmy se řadí:

- genetické algoritmy
- evoluční programování
- evoluční strategie
- genetické programování

Bayesovská klasifikace

Metody bayesovské klasifikace patří mezi ty složitější a ne tak často používané. Opírají se o Bayesovy věty o podmíněných pravděpodobnostech a jsou studovány v souvislostech se strojovým učením. Uplatnění pak metody bayesovské klasifikace nacházejí právě v systémech pro dobývání znalostí.

Metody založené na analogii

Jak název výstižně napovídá, do této podkapitoly nespadá pouze jedna metoda, nýbrž několik. Níže vyjmenované metody spojuje princip, který je založený na analogii, čili pokud narazíme na dosud neznámou situaci, použijeme řešení, které se už dříve osvědčilo v podobném případě.

Mezi metody založené na analogii patří:

- shlukování
- případové usuzování
- učení založené na instancích
- líné učení
- paměťové učení
- pravidlo nejbližšího souseda

1.5 Nástroje pro data mining

Nejrůznějších nástrojů pro dolování dat je na trhu nemalé množství. S přihlédnutím k tomuto faktu byla pro bakalářskou práci vzhledem k její povaze zvolena pouze část využívaných nástrojů, a to ty nejznámější a nejčastěji využívané.

Nástroje pro data mining by se daly (jako ostatně téměř všechny typy softwarových programů) rozdělit na dvě hlavní části, a to:

- Na nástroje nabízené komerčně
- Na nástroje dostupné jako open source

Mezi nejčastěji komerčně nabízení nástroje data miningu patří:

- IBM SPSS Modeler
- SAS® Data Mining
- Oracle Data Mining
- Microsoft SQL

Nejvíce používanými open source nástroji pak jsou:

- RapidMiner
- KNIME
- Orange
- R

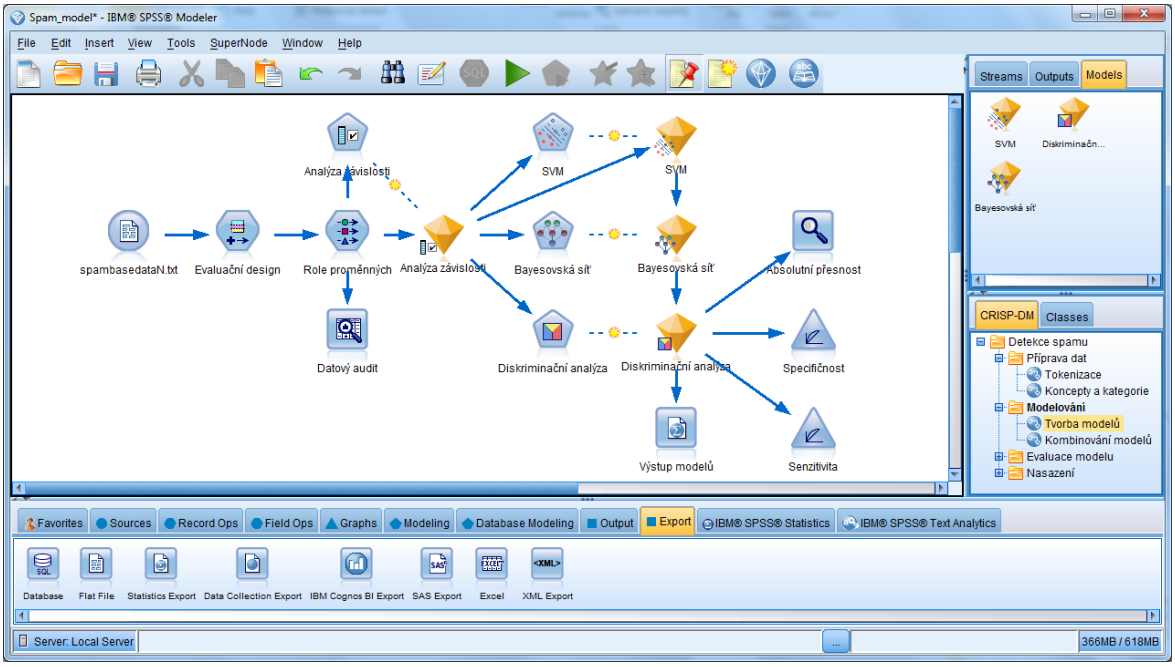
1.5.1 Charakteristiky vybraných nástrojů

V této podkapitole budou nyní stručně popsány vlastnosti jednotlivých výše vyjmenovaných data miningových nástrojů.

Komerčně nabízené nástroje

Program IBM SPSS Modeler, jak už název napovídá, vyvinula společnost IBM, která je obecně známá kvalitními produkty v nejrůznějších oblastech IT a ani v oblasti data miningu tomu není jinak.

IBM SPSS Modeler, jehož uživatelské rozhraní můžeme vidět na *obrázku č. 4 Rozhraní nástroje SPSS Modeler* , je nástroj, který je na trhu poskytován ve třech základních verzích – Professional, Premium a Server.

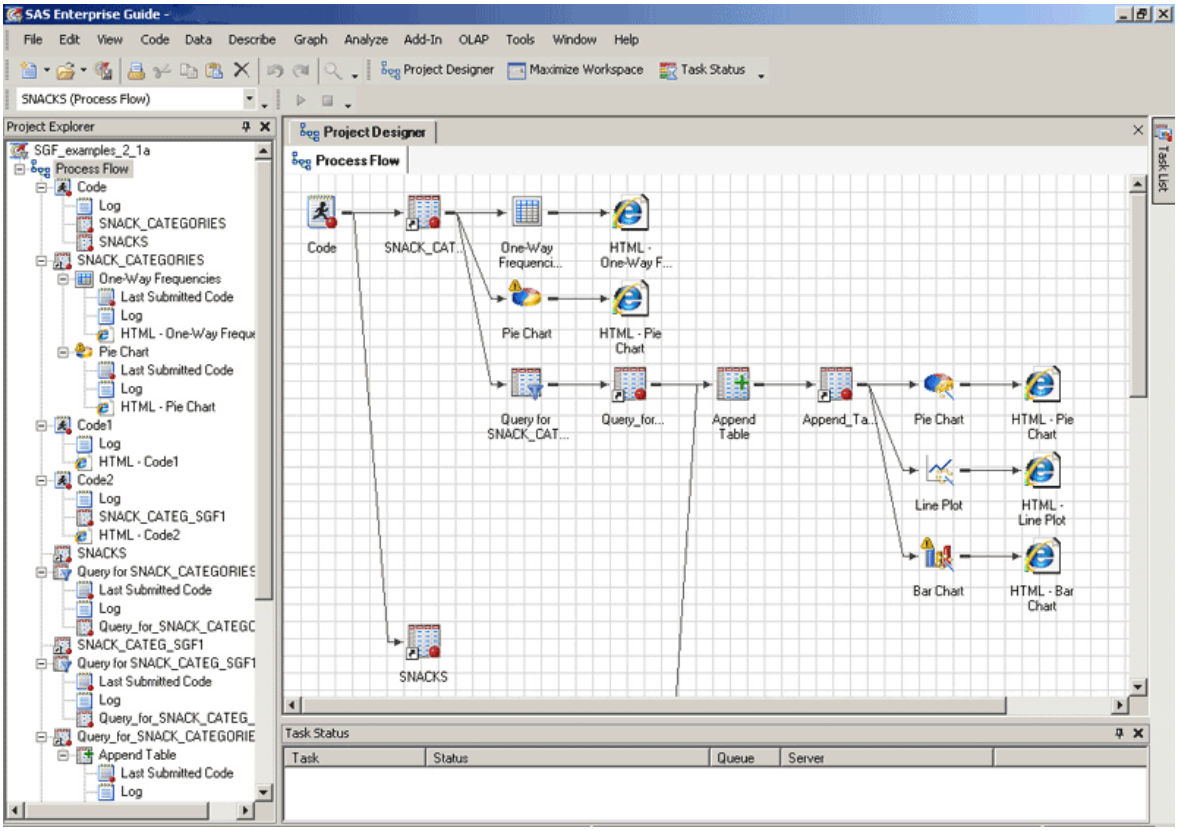


Obrázek 4 Rozhraní nástroje SPSS Modeler
Zdroj: http://www.acrea.cz/upload/img/ibm_spss_modeler_15_printscreen.png

Jednotlivé verze se dle [9] od sebe liší v různých nadstavbách. Zatímco Professional slouží hlavně pro predikce ze strukturovaných číselných dat, verze Premium se zaměřuje na nestrukturovaná data získaná převážně z aktivit v internetovém prostředí (weby, e-maily, zpětná vazba…). Verze Server pak oproti předchozím produktům umožňuje sdílení výsledků hledání a usnadňuje tak týmovou práci na rozsáhlejších manažerských projektech.

Nástroj SAS, jehož uživatelské rozhraní můžeme vidět na *obrázku č. 5 Rozhraní nástroje SAS*, je ve svých základních funkcích víceméně srovnatelný s IBM SPSS Modelerem. Dle

oficiálních stránek společnosti SAS [10] je produkt poskytovaný v oblasti data miningu postaven hlavně na metodách asociačních pravidel a dále pak na segmentaci a sekvenční analýze.



Obrázek 5 Rozhraní nástroje SAS
Zdroj: http://www.sas.com/resources/screenshot/soda_eg.gif

Dalším často užívaným nástrojem je Oracle Data Mining, který dle [12] poskytuje výkonné algoritmy dolování dat. Posledním výše zmíněným je pak produkt Microsoft SQL, který i když se specializuje hlavně na běžné SQL dotazy, obsahuje v sobě i rozšíření pro data mining.

Open source nástroje

Program RapidMiner (známý rovněž pod svým starším názvem YALE) je zaměřený především na dolování dat z oblasti finančnictví a marketingu. Tím se liší od druhého zmiňovaného nástroje KNIME.

KNIME je dle [11] hojně využívaný v oblasti biochemie, neboť jeho součástí jsou speciální rozšíření, která jsou právě pro toto odvětví potřebná.

Dalším nástrojem jmenovaným výše je program Orange. Ten se oproti předchozím dvěma vyznačuje velice jednoduchým a uživatelsky přívětivým rozhraním, které opět dle [11] nezatěžuje uživatele přítomností mnoha zřídkakdy používaných funkcí.

Posledním z velice často používaných, volně dostupných nástrojů je program R, který se zaměřuje hlavně na statistické metody z oblasti data miningu. Oproti předchozím třem se vyznačuje tím, že (pokud uživatel nemá nainstalované některé z přídavných package) se ovládá pomocí příkazového řádku.

Obecně vzato velkou výhodou všech open source nástrojů je velikost komunity, která je užívá. Díky tomu na internetu nenalezneme pouze oficiální tutoriály a nápovědy, ale i spoustu uživatelských fór a videí, na kterých lze najít překvapivé množství rad, tipů a triků, jak si práci usnadnit, a tím ji zároveň i zefektivnit.

1.5.2 Výběr nástroje

Konečný výběr nástroje samozřejmě závisí na mnoha faktorech. Pokud bychom měli jmenovat některé z nich, budou to v prvé řadě:

- Schopnost nástroje vykonat požadované úkoly
- Zkušenosti s programem
- Technická podpora
- Firemní know-how
- Cena licence (v případě komerčně nabízených data miningových nástrojů)

Co se týče výběru konkrétních nástrojů pro účely projektu bakalářské práce, ten byl začleněn do praktické části, a sice do podkapitoly *2.3.2 Příprava dat – Technologické nástroje*.

2 PRAKTICKÁ ČÁST

Praktická část bakalářské práce má za cíl ukázat konkrétní postupy při data miningových úlohách, a to, jak již z názvu bakalářské práce vyplývá, na datech zdravotní pojišťovny.

Datová analýza byla prováděná coby reálná zakázka, kterou společnost KOMIX s.r.o. na základě předložené nabídky získala.

Vzhledem k citlivosti údajů nebude v praktické části bakalářské práce použito jméno vybrané zdravotní pojišťovny coby zákazníka ani názvy jednotlivých zdravotnických zařízení, kterých se výstupy analýz týkají.

2.1 Využití data miningových úloh pro manažerské rozhodování ve zdravotní pojišťovně

V této kapitole se budeme zabývat výběrem vhodných metod pro naši aplikaci z jednotlivých data miningových metod popsaných v předchozí části.

2.1.1 Nejpoužívanější data miningové metody

Obecně pro manažerské rozhodování jsou nejpoužívanějšími metodami rozhodovací stromy a asociační pravidla v podobě tzv. analýzy nákupního košíku. Tyto dvě metody mohou být aplikovány i na datech pro potřeby zdravotní pojišťovny, a to následovně:

Rozhodovací stromy pro abnormální výkyvy za proplácené léky či výkony. V obdobných, zkušenostmi prověřených příkladech pak jednotlivé uzly signalizují podezřelé anomálie, čímž jasně ukazují, kam by měl manažer při svém rozhodování směřovat svou pozornost.

Analýza nákupního košíku vypovídá o tom, co nejčastěji člověk nakoupí spolu s nějakým jiným produktem. Tato metoda se dá pro potřeby zdravotní pojišťovny aplikovat na datech z lékařských předpisů, a to asi takto: *Pokud pacientovi byl předepsán lék A, byl mu nejspíš předepsán i lék B.* Stane-li se vše podle daného scénáře, je vše v pořádku a nejsou třeba

žádné následné kontroly. Kdyby ovšem z ničeho nic jeden lékař začal pravidelně předepisovat místo léku B léčebný přípravek C, už by mohlo jít o pokus o podvod pojišťovny. Takovým příkladem je, když se některý lékař dohodne s určitým distributorem, že namísto běžného přípravku bude za jistou provizi předepisovat pro daný úkon dražší lék. V praxi se jedná o dost častý případ, který se ale při pravidelných validacích, které v informačních systémech pojišťoven probíhají denně, špatně odhalují, a proto je potřeba právě pokročilých datových analýz jako je tato.

Dalším příkladem, jak by se dala analýza nákupního košíku využít, je odhalení nepředepisování generik, čili ekvivalentních léků k originálním léčebným přípravkům. Jde vlastně o speciální případ výše popisovaného problému, kdy lékaři předepisují originální (tedy dražší) léky namísto přípravků se stejnou léčebnou látkou (generikum), které jsou výrazně levnější. Pokud se manažer zdravotní pojišťovny díky analýze nákupního košíku o podobném jednání dozví, může zasáhnout a vyjednat s konkrétním lékařem, aby začal generika předepisovat a zdravotní pojišťovna tak ušetřila nemalé výdaje.

2.1.2 Metody použité pro vybranou zdravotní pojišťovnu

Nejčastější metody popsané v předchozí kapitole nakonec pro zakázku od vybrané zdravotní pojišťovny vybrány nebyly. Důvodem k tomu byl především požadavek zákazníka, jehož představa byla následující:

Ve zdravotní pojišťovně byl víceméně náhodou odhalen jeden případ, kdy lékař neoprávněně vykazoval výkon zasádrování zlomeniny. Pracovně byl tento podvod nazván *„zasádroval celou vesnici“* a šlo ve své podstatě o toto – lékař provedl tolik výkonů zasádrování končetin, že kdyby to měla být pravda, musel by vzhledem k procentuální propojištěnosti v dané lokalitě doslova zasádrovat každého člověka ve vesnici.

Klienta v rámci vypsaného projektu zajímalo, kolik obdobných, dosud nezjištěných případů se po celé České republice ještě vyskytuje. Proto bylo třeba odrazit se od doposud nezavedených a nevyzkoušených metod a přijít s něčím novým, vytvořeným na míru potřebám zákazníka.

Díky velice přesné představě zákazníka, bylo jasné z čeho vycházet. Stěžejním údajem se stala zmiňovaná *propojištěnost*, čili údaj kolik procent obyvatel je ve vybraném místě (rozděleno na okresy dle číselníků pojišťovny) klienty vybrané zdravotní pojišťovny. S tou se pak pracovalo v modelech založených na metodě regresní analýzy, která byla popsána výše v teoretické části.

Další použité modely, které byly na data od vybrané zdravotní pojišťovny aplikovány, pak vycházely rovněž ze statistiky podobně jako regresní analýza. Jejich princip spočíval na rozdělení dostupných dat na percentily, a i když tato metoda (bližší popis na konkrétních modelech v podkapitole *2.4.2 Modely*) nepatří mezi žádné z výše popsaných v teoretické části bakalářské práce, osvědčila se při hledání abnormalit v tak velkém množství dat.

2.2 Předmět plnění zakázky

Zdravotní pojišťovny se velice často potýkají s podvodným vykazováním lékařských výkonů. Pro zachycení běžných nesrovnalostí mají jejich systémy nastaveny celou řadu kontrol: od prerevizí, revizí, až po validace.

Výše zmíněné kontroly ale bohužel nezachytí veškeré podvodné aktivity, jak již bylo uvedeno výše v podkapitole *2.1.1 Nejpoužívanější data miningové metody*. Proto se daná zdravotní pojišťovna rozhodla v rámci dlouhodobých úspor investovat své peníze právě na analýzu dat, jejíž účel je identifikace a predikce případů nestandardního, a tedy potenciálně neoprávněného čerpání nákladů na zdravotní péči, a návrh příslušných algoritmů, které by vedly k určení těchto případů.

Cílem firmy KOMIX s.r.o. jako dodavatele data miningové služby pak bylo dodat seznam zdravotnických zařízení, která dle výsledků analýz čerpala náklady za vykázané výkony nestandardním způsobem a která jsou tudíž podezřelá z neoprávněné činnosti.

2.3 Získání a příprava dat

Základem pro každou analýzu jsou především data a správný pohled na ně. Samotná analýza není až tak složitá, pokud jsou data pro jednotlivé části pečlivě a dobře připravena. I z toho hlediska je proto výhodné věnovat této fázi dostatečnou pozornost a nezanedbat ji, neboť se do budoucna může jednat o značnou úsporu drahocenného času.

2.3.1 Získání dat

Data potřebná pro analýzu byla získána přímo ze zdravotní pojišťovny dle podmínek stanovených na samém počátku před podepsáním smlouvy, ve které jsou písemně zahrnuty.

Vybraná zdravotní pojišťovna coby objednatel poskytla pro analýzu následující data:

- Položky s vykázanou a uznanou zdravotní péčí (koruny, body)
- Položky preskripcí/úhrad za léky a za prostředky zdravotní techniky (PZT)
- Statistické údaje o počtech obyvatel
- Statistické údaje o tzv. propojištěnosti
- Údaje o závažných lékových interakcích
- Statistické zařazení klientů do interních skupin dle klasifikace vybrané zdravotní pojišťovny
- Další údaje, které by mohly být pro účely plnění zakázky relevantní

Pro analýzu dat nebyly poskytnuty žádné osobní údaje pojištěnců. Modely (viz kapitola 2.3.2. Modely níže), kde se vyskytl poměr unikátních rodných čísel k celkovému počtu výskytu rodných čísel, respektive celkovému počtu výkonů, byly odvozeny od čísla pojištěnce, které splňuje kritérium unikátnosti stejně jako rodné číslo.

2.3.2 Příprava dat

Hlavní součástí přípravné fáze byla nutnost zorientovat se v datech, která zdravotní pojišťovna pro analýzu poskytla.
V této části projektu z hlediska dodavatele vždy záleží především na komunikaci s klientem, přičemž zájem pochopit správně vstupy pro analýzu je, respektive by měl být, oboustranný.
Objednavatel si musí být dopředu vědom faktu, že nejpodstatnější pro následné manažerské rozhodování je, aby analýzy byly skutečně co nejkvalitnější, a tím pádem vedly k co nejstrategičtějším budoucím krokům. A to mu může zajistit pouze dobře informovaný analytik.

Číselníky

Jednou z částí přípravy dat je vytvoření tzv. číselníků, tedy pomocných tabulek pro následnou práci s velkoobjemovými databázemi. Číselníky jsou nezbytné, protože zmiňované databáze obsahují velké množství sloupců, v kterých se zpravidla vyskytují pouze kódy a zkratky, které by byly obtížně uchopitelné pro konečnou interpretaci výsledků, ale i pro samotný náhled na data.

Příklady číselníků použitých pro projekt vybrané zdravotní pojišťovny:

- ccdiag – číselník s diagnózami
  - obsahuje sloupce: id (kód diagnózy), nazev (název diagnózy), pohlavi (pohlaví Z nebo M v případě, že se jedná o diagnózu, která se vyskytuje pouze u daného pohlaví)
- ccdruhvyk – číselník s druhy výkonů
  - obsahuje sloupce: id (kód druhu výkonu), nazev (název druhu výkonu)
- ccdruhzz – číselník s druhy zdravotnických zařízení
  - obsahuje sloupce: id (kód přiřazený danému zdravotnickému zařízení), nazev (název zdravotnického zařízení)
- ccodb – číselník odborností pracovišť
  - obsahuje sloupce: id (číslo přiřazené dané odbornosti), odb_kod (kód odbornosti), nazev (název odbornosti)

- ccstatus_schv – číselník vyplývající z revizí systému pojišťovny
  - obsahuje sloupce: id (číslo daného nálezu systému), nazev (název zjištěného nálezu)
- ccokres – číselník s okresy
  - obsahuje sloupce: id1, id2 (identifikační čísla okresů), nazev (název okresu), kraj (číslo kraje)
- ccokresobyv – jeden z nejdůležitějších číselníků pro projekt, který obsahuje počty obyvatel po jednotlivých letech a daných okresech
  - obsahuje sloupce: rok (roky 2009-2012), okres_id2 (číslo okresu id2 z předešlého číselníku), obyv_muzu (počet muž v daném okrese), obyv_zen (počet žen v daném okrese), obyv (celkový počet obyvatel v daném okrese)
- ccokrespoj – další z velmi důležitých číselníků pro projekt zahrnující údaje o pojištěncích vybrané zdravotní pojišťovny po jednotlivých letech a po daných okresech
  - obsahuje sloupce: rok (roky 2009-2012), okres_id2 (číslo okresu), pojist_muzu (počet mužských klientů vybrané pojišťovny), pojist_zen (počet klientek vybrané pojišťovny), pojist (počet pojištěnců dané zdravotní pojišťovny)
- ccpoj – nejobsáhlejší číselník co do počtu řádků zahrnující obecné informace o klientech vybrané zdravotní pojišťovny
  - obsahuje sloupce: id (číslo pojištěnce), pohlavi (pohlaví M/Z), okres (číslo okresu), vek (věk pojištěné osoby)

Souvislosti

Hledání souvislostí a spojitostí mezi daty je něco, co se velice těžko popisuje, nicméně i to je část, která patří do přípravné fáze samotné analýzy.

Pro analyzování dat je zapotřebí mít nápady na to, co by mohlo ukázat v tomto konkrétním případě na podvodné čerpání nákladů na zdravotní péči, a hlavně vědět, jak danou myšlenku realizovat pomocí dostupných datových zdrojů a technologických nástrojů.

Pro tuto část procesu jsou nesmírně důležité hlavně zkušenosti analytika, neboť se rozhodně nejedná o triviální záležitost, jak by se mohlo na první pohled zdát. Ne všechna

dostupná data mají vypovídající hodnotu pro daný problém. Úkolem analytika v této fázi je vybrat k odpovídajícím datům vhodné metody, které se budou při následném analyzování používat.

Technologické nástroje

S volbou vhodných dat a metod přichází i rozhodování jaké aplikační nástroje pro analýzu použít.

Nejpoužívanější nástroje byly zmiňovány a podrobněji rozebírány již v teoretické části, a to v kapitole *1.5 Nástroje pro data mining*.

Pro práci na projektu zdravotní pojišťovny byly zamítnuty komerční data miningové nástroje, neboť se jednalo o menší zakázku, a tudíž by se do jejich koupě firmě nevyplatilo investovat. Proto se vybíralo pouze z open source nástrojů.

Požadavky pro zhotovení projektu splňovalo více popisovaných nástrojů, z nichž po vyzkoušení vyšel nejlépe program R, a to hlavně díky tomu, jak dobře a intuitivně se s ním pracuje.

Jako doplňující program pro přípravu dat, která do analýzy vstupují, pak byl zvolen databázový program MS SQL Server Management Studio a některé číselníky byly vygenerované v aplikaci QlikView, kterou vyvinula firma KOMIX s.r.o.

2.4 Analýza dat

Analýza je část, v níž dochází k samotnému zpracovávání dat, a zpravidla probíhá v několika fázích. Ty na sebe navazují, respektive se navzájem doplňují. Díky již vzniklým výstupům se rodí nové nápady a možnosti, upravují se skripty a výsledky se znovu přehodnocují a předělávají, aby co nejpřesněji odpovídaly vymezenému zadání úkolu.

2.4.1 Analyzovaná data

Pojišťovna pro analýzu poskytla údaje o výkonech za necelé čtyři roky (rok 2009, 2010, 2011 a zhruba šest měsíců za rok 2012), což bylo více jak dvě stě miliónů řádků.

Pro samotnou datovou analýzu, s ohledem na důležitost aktuálnosti údajů, se nakonec použily řádky výkonů vykázaných za období 1. 7. 2011 – 30. 6. 2012.

Po dohodě se zdravotní pojišťovnou se také upravil seznam analyzovaných zdravotnických zařízení, a to dle odborností pracovišť. Do analýzy byla vybrána zařízení s těmi kódy odborností, ve kterých se ročně vykáže nejvíce výkonů, a tudíž se na celkových výdajích zdravotní pojišťovny podílí největší měrou.

Pro kontroly byla vybrána zařízení následujících odborností:

- Samostatná ordinace lékaře specialisty
- Samostatná ordinace praktického lékaře - stomatologa
- Samostatná ordinace praktického lékaře - gynekologa
- Samostatná ordinace praktického lékaře pro dospělé
- Samostatná ordinace praktického lékaře pro děti a dorost
- Domácí ošetřovatelská péče/home-care

Do kontrol nebyla zařazena zdravotnická zařízení, u kterých byl:

- Počet unikátních RČ v první polovině sledovaného období < 10
- Datum ukončení činnosti zařízení leželo ve sledovaném období

2.4.2 Modely

Pro datovou analýzu bylo vybráno celkem pět modelů, které budou podrobně popsány v této podkapitole.

Všechny modely obsahují společný prvek, a sice číslo IPD (index potenciálního darebáctví), jehož pracovní název se i zdravotní pojišťovně coby objednateli služby natolik líbil, že se uchytil a přetrval až do konce projektu.

Index potenciálního darebáctví slouží jako prostředek pro interpretaci výsledků. Dílčí hodnoty IPD z jednotlivých modelů jsou sčítány a výsledná výše indexu je pak přímo úměrná výši pravděpodobnosti, že dané zdravotnické zařízení v uplynulém období čerpalo neoprávněně úhrady za výkony a materiál.

IPD model č. 1 – Netypický poměr unikátních RČ/ celkový počet RČ ve všech výkonech

Model netypického poměru unikátních rodných čísel k celkovému počtu rodných čísel ve všech výkonech má za cíl poukázat na podezřelé vykazování výkonů podle jednotlivých kódů odbornosti, tj. například na nestandardní chování jednoho praktického lékaře pro dospělé mezi všemi ostatními praktickými lékaři pro dospělé.

Výpočet IPD pro tento model vychází z percentilů pro každou jednu odbornost (parametr p_1). Odtud se vezme percentil 50%, tedy medián (parametr p_2), který je posléze porovnávám s mediánem jednotlivých zařízení (parametr p_3).

Vzorec pro výpočet IPD z modelu č. 1 vypadá pak následovně:

(2) $IPD_1 = 100 * abs (p_2 – p_3)$,

přičemž text generovaného výstupu obecně zní:

+ IPD_1 z modelu 1: Celkový poměr URČ/RČ je netypický. Medián pro odbornost p_1 je p_2, toto zařízení má p_3.

Názorný příklad jednoho náhodně vybraného výstupu z datové analýzy vypadá takto: *+34 z modelu 1: Celkový poměr URČ/RČ je netypický. Medián pro odbornost 001 je 0.29, toto zařízení má 0.63*.

Vygenerovaný text bychom pak interpretovali následujícím způsobem:

Dané zdravotnické zařízení obdrželo na základě výpočtu třicet čtyři trestných bodů z modelu 1 s hlášením: Celkový poměr URČ/RČ je netypický. Medián pro odbornost 001 - Pracoviště praktického lékaře pro dospělé – je 29%, ale dané zařízení dosáhlo 63%. Obecně to znamená, že pokud k praktickému lékaři pro dospělé přijde pacient, s dvaceti devíti procentní pravděpodobností se bude jednat o pacienta, který v daném období u lékaře ještě nebyl. V daném zařízení ale procento unikátních pacientů bylo více jak dvojnásobné – šedesát tři procenta všech pacientů byla u svého praktika pouze jednou.

V praxi to indikuje kupříkladu to, že podezřelé zařízení by mohlo neoprávněné vykazovat výkony na náhodné pacienty, kteří u lékaře vůbec nebyli.

IPD model č. 2 – Netypický poměr unikátních RČ/ celkový počet RČ pro daný výkon

Model netypického poměru unikátních rodných čísel k celkovému počtu rodných čísel pro daný výkon je určen k odhalení neoprávněného čerpání nákladů za jednu konkrétní činnost.

Hodnota IPD pro tento model vychází opět z percentilů. Medián (parametr p_3) je vypočítaný pro každý jeden kód výkonu a porovnává se s mediánem stejného výkonu (parametr p_4) konkrétních zdravotnických zařízení stejné kategorie.

Vzorec pro výpočet IPD z modelu č. 2 vypadá pak následovně:

(3) $IPD_2 = 50 * abs (p_3 - p_4)$,

přičemž text generovaného výstupu obecně zní:

+ IPD_2 z modelu 2: Poměr URČ/RČ pro výkony p_1 (p_2) je netypický. Medián p_3, hodnota tohoto zařízení p_4, vykázalo p_5 výkonů,

kde:

- parametr p_1 je kód daného výkonu
- parametr p_2 je název výkonu
- parametr p_3 je hodnota mediánu pro daný výkon (viz popis výše)
- parametr p_4 je aktuální hodnota mediánu pro zařízení (viz popis výše)

- parametr p_5 je počet vykázaných výkonů tohoto druhu.

Názorný příklad jednoho náhodně vybraného výstupu z datové analýzy vypadá takto: *+32 z modelu 2: Poměr URČ/RČ pro výkony 51875 (PŘILOŽENÍ MĚKKÉHO OBVAZU (ZINKOKLIH, ŠKROBOVÝ OBVAZ) NA DOLNÍ NEBO HORNÍ KONČETINU) je netypický. Medián je 0.89, toto zařízení má 0.26. Počet výkonů je 410.*

Vygenerovaný text bychom pak interpretovali následujícím způsobem:

Dané zdravotnické zařízení obdrželo na základě výpočtu třicet dva trestných bodů z modelu 2 s hlášením: Poměr URČ/RČ pro výkon s kódem 51875 – přiložení měkkého obvazu na dolní nebo horní končetinu je netypický. Medián daného výkonu představuje 89 % pravděpodobnost, že ošetřovaný pacient bude unikátní, do daného zařízení ale přichází pouze 26% unikátních pacientů a zbytku je tento výkon za sledované období prováděn opakovaně. Počet provedených výkonů tohoto druhu je 410.

Pohled do praxe ukazuje, že pro vybraný výkon je podle jeho percentilového rozložení typické pouze jedno ošetření, tudíž je velká pravděpodobnost, že dané zařízení provádí tento úkon opakovaně, aniž by byl vyloženě nutný. Tím toto pracoviště způsobuje zdravotní pojišťovně zbytečné výdaje navíc.

IPD model č. 3 – Netypický počet konkrétních výkonů

Model netypického počtu konkrétních výkonů je navržen ke hledání a odhalení podvodného vykazování vybraných úkonů.

Třetí model je společně se čtvrtým, co se přípravy a zpracování dat týče, ze všech pěti modelů vůbec tím nejnáročnějším a nejsložitějším.

Součástí modelu[2] Netypického počtu konkrétních výkonů jsou regresní rovnice, které byly vytvořeny zvlášť pro každý typ výkonu a které obsahují jeden nebo více z následujících parametrů:

- R_1_PocetAktZarizeniOdbornostiVOkrese - 1/počet aktivních zařízení v okrese se stejnou odborností (mající alespoň 1 doklad ve sledovaném období),
- ObyvatelVOkrese – počet obyvatel v okrese zařízení,
- PojistencuVOkrese – počet pojištěnců v okrese zařízení,
- PropojistenostVOkrese – poměr PojistencuVOkrese / ObyvatelVOkrese,
- PocetDokladuCelkem – počet dokladů zařízení celkem,
- PocetVykonuCelkem – počet výkonů zařízení celkem,
- R_PocetVykonu_PocetDokladuCelkem – poměr PocetVykonuCelkem / PocetDokladuCelkem,
- PocetUnikRCVykonuCelkem – počet unikátních RČ u daného výkonu,
- R_PocetUnikRC_PocetVykonuCelkem – poměr počtu unikátních RČ a počtu výkonů.

Hodnota indexu potenciálního darebáctví pro daný model vychází z poměru dvou parametrů, a to aktuálního počtu výkonů (parametr p_4) děleného očekávaným počtem výkonů (parametr p_3).

Vzorec pro výpočet IPD z modelu č. 3 vypadá tedy následovně:

(4) $IPD_3 = \frac{p4}{p3}$,

přičemž text generovaného výstupu obecně zní:

+ IPD_3 z modelu 3: Počet výkonu p_1 (p_2) je netypicky vysoký. Očekáváno p_3,vykázáno p_4.

kde:

- parametr p_1 je kód daného výkonu
- parametr p_2 je název výkonu

[2] Pro následný výpočet IPD byly použity pouze ty regresní modely, jejichž R^2 bylo >= 0,8. Do výsledků se dále nezapočítávala zařízení, u kterých byl poměr p_4 / p_3 <= 1.5 (viz vzorec níže)

- parametr p_3 je očekávaná výše počtu vykázaných výkonu dle regresní rovnice
- parametr p_4 je počet vykázaných výkonů tohoto druhu

Názorný příklad jednoho náhodně vybraného výstupu z datové analýzy vypadá takto: *+35 z modelu 3: Počet výkonů 935 (SUBGINGIVÁLNÍ OŠETŘENÍ) je netypicky vysoký. Očekáváno 5, vykázáno 176.*

Vygenerovaný text bychom pak interpretovali následujícím způsobem:

Zdravotnické zařízení obdrželo na základě výpočtu IPD pro model č. 3 třicet pět trestných bodů s hlášením: Počet výkonů mající kód 935 – subgingiválního ošetření je netypicky vysoký. Bylo očekáváno pět těchto výkonů, zařízení však vykázalo sto sedmdesát šest výkonů.

V reálu toto hlášení může být znakem toho, že zařízení neoprávněně vykazuje určitý druh výkonu, aniž by ho skutečně vykonávalo, nebo kupříkladu to, že si zve pacienty zvláště na tento úkon. Důvodem by ovšem mohlo být i to, že se dané zdravotnické zařízení prostě jen specializuje na tuto konkrétní činnost, a proto ji logicky vykazuje častěji, než s čím regresní rovnice počítá.

IPD model č. 4 – Nečekaně vysoká cena za doklady v 2. polovině sledovaného období

Model nečekaně vysoké ceny za doklady v druhé polovině sledovaného období se zakládá na srovnání dvou po sobě jdoucích období mezi sebou.

Čtvrtý model je podobně jako ten předešlý založen na regresní analýze. Regresní rovnice pro proměnnou CenaDokladu_2P (cena dokladů v druhé polovině sledovaného období) je vytvořena zvlášť pro každý typ odbornosti a zahrnuje jeden nebo více prediktorů z následujícího seznamu:

- CenaDokladu_1P – cena všech dokladů v první polovině sledovaného období
- PocetUnikRCVykonu_1P – počet unikátních rodných čísel u výkonů v první polovině sledovaného období

- PocetUnikRCVykonu_2P – počet unikátních rodných čísel u výkonů v první polovině sledovaného období
- PocetNovychUnikRC_2P – počet nových unikátních rodných čísel v druhé polovině sledovaného období
- R_PocetNovychUnikRC_2P_1P – poměr počtu nových unikátních rodných čísel v obou polovinách sledovaného období
- PocetUnikRCVykonuCelkem – počet unikátních RČ u daného výkonu,
- R_PocetUnikRC_PocetVykonuCelkem – poměr počtu unikátních RČ a počtu výkonů.

Výsledná hodnota IPD je vypočítaná z poměru dvou parametrů, a sice ze skutečné hodnoty nárokované ceny všech dokladů v druhé polovině sledovaného období (parametr p_2) krácené očekávanou hodnotou nárokované ceny za stejný časový úsek (parametr p_1).

Vzorec pro výpočet IPD z modelu č. 4 má následující podobu:

(5) $IPD_4 = \frac{p2}{p1}$,

přičemž vygenerovaný text výstupu obecně zní:

+ IPD_4 z modelu 4: Cena dokladů v 2. polovině období je nečekaně vysoká. Očekáváno p_1 Kč, zařízení vykázalo p_2 Kč. V 1. pol. vykázáno p_3 Kč,

kde poslední nepopsaný parametr p_3 představuje vykázanou cenu dokladů v 1. polovině období.

Názorný příklad jednoho náhodně vybraného výstupu z datové analýzy vypadá takto: *+2 z modelu 4: Cena dokladů v 2. polovině období je nečekaně vysoká. Očekáváno 89698 Kč, zařízení vykázalo 159888 Kč. V 1. pol. vykázáno 37204 Kč.*

Automaticky generovaný text bychom pak interpretovali následujícím způsobem:

Zařízení obdrželo dva body IPD ze čtvrtého modelu: Cena dokladů v 2. polovině období byla nečekaně vysoká, neboť se očekávalo 89 698 Kč, ale zařízení vykázalo 159 888 Kč,

tedy částku téměř dvojnásobnou. V 1. polovině období pak zařízení vykázalo částku 37 204 Kč.

V praxi tento jev může opět poukazovat na možné neoprávněné vykazování náhrad.

IPD model č. 5 – Velká změna počtu nových RČ v 2. polovině období

Poslední model Velká změna počtu nových RČ v druhé polovině období je modelem, kde se kontroluje procento nových RČ pacientů, kteří byli ošetřeni v druhé polovině sledovaného období a přitom nebyli ošetřeni v první polovině sledovaného období. Tento údaj se sleduje ve srovnání se zařízeními stejné odbornosti.

Hodnota IPD nabývá pro tento model pouze dvou hodnot (jedničky, nebo dvojky), a to podle toho, kde se nachází medián pro zařízení stejné odbornosti.

- Pokud je hodnota mediánu pro zařízení stejné odbornosti menší rovna hodnotě mediánu daného zařízení (parametr p_3) pro 1 % percentil (či větší rovna hodnotě pro 99 % precentil), pak $IPD_5 = 2$.
- Pokud je pak hodnota mediánu pro zařízení stejné odbornosti menší rovna hodnotě mediánu daného zařízení (parametr p_3) pro 5 % percentil (či větší rovna hodnotě pro 95 % precentil), dílčí hodnota $IPD_5 = 1$.

Obecně pak vzorce pro určení této hodnoty můžeme vyjádřit takto:

(6) Když p_3 je <= percentil 1 nebo >= percentil 99, pak $IPD_5=2$.
(7) Když p_3 je <= percentil 5 nebo >= percentil 95, pak $IPD_5=1$.

přičemž generovaný text výstupu zní:

+ IPD_5 z modelu 5: Podíl nových URČ v 2. pol. období oproti první polovině je netypický. Medián pro odbornost p_1 je p_2, zařízení má p_3,

kde:

- parametr p_1 představuje kód odbornosti zařízení
- parametr p_2 představuje procento nových RČ - medián pro všechna zařízení stejné odbornosti

- parametr p_3 představuje procento nových RČ – aktuální hodnota pro zařízení (viz výše)

Názorný příklad jednoho náhodně vybraného výstupu z datové analýzy vypadá takto: *+1 z modelu 5: Podíl nových URČ v 2. pol. období oproti první polovině je netypický. Medián pro odbornost 002 je 0.27, zařízení má 0.59.*

Automaticky generovaný text bychom pak interpretovali následujícím způsobem:

Dané zdravotnické zařízení obdrželo na základě výpočtu jeden trestný bod z modelu č.5 s hlášením: Podíl nových URČ v 2. pol. období oproti první polovině je netypický. Medián pro odbornost 002 – Pracoviště praktického lékaře pro děti a dorost – je 27%, ale dané zařízení dosáhlo 59%. Obecně to znamená, že je 27 % pravděpodobnost, že k lékaři pro děti a dorost přijde nový pacient, kdežto v daném zařízení je tato pravděpodobnost více jak dvojnásobná.

V praxi to může opět znamenat, že podezřelé pracoviště neoprávněně pobírá náhrady tak, že vykazuje úkony na pacienty, kteří u lékaře nebyli.

2.5 Výsledky datové analýzy a zhodnocení přínosů bakalářské práce

Jak už název kapitoly napovídá, v této části práce budou rozebrány výsledky provedené datové analýzy a poté zhodnoceny přínosy bakalářské práce.

2.5.1 Výsledky datové analýzy

V této podkapitole budou nyní rozebrány výsledky datové analýzy, a to v obecné rovině, neboť, jak již bylo zmíněno na samém počátku praktické části, jednotlivé modely pracovaly s citlivými údaji, které zdravotní pojišťovna o zdravotnických zařízeních

poskytla pro účely datové analýzy, a proto v rámci bakalářské práce nesmí být použit konkrétní výstup, tedy tabulka, která se po dokončení projektu předávala klientovi.

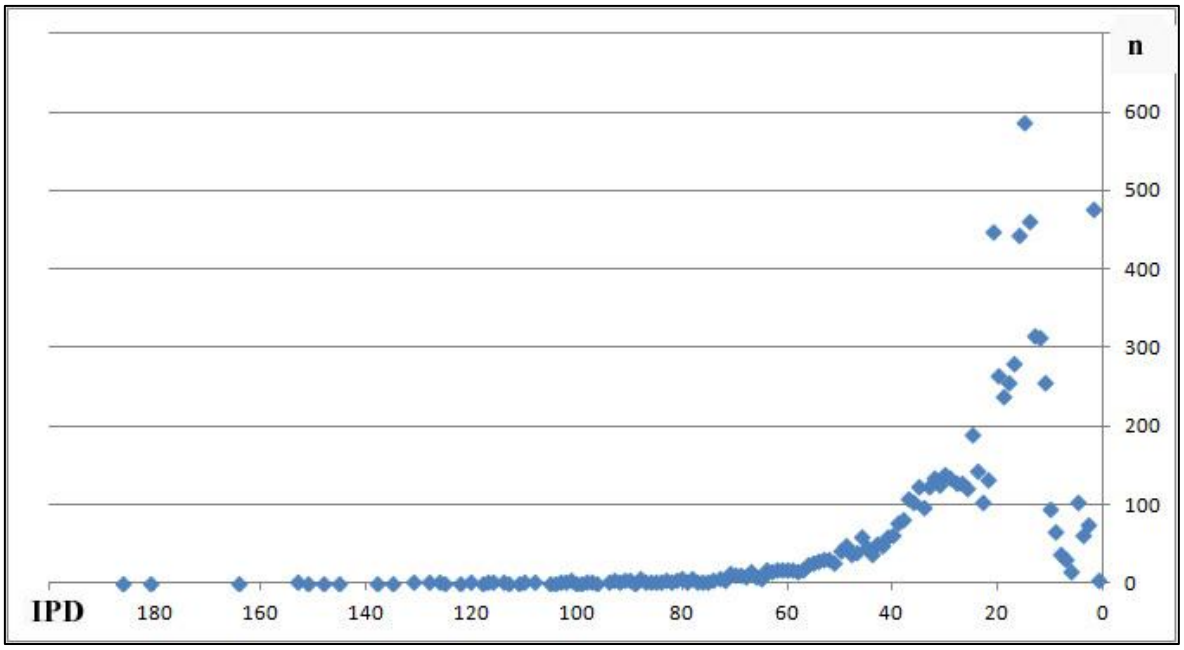
Celkem do analýzy vstupovalo okolo 20 000 zařízení vybraných na základě kritérií popisovaných v kapitole *2.4.1 Analyzovaná data*, přičemž při výstupu z analýzy vyšlo s podezřením na podvodné vykazování 8 108 z nich.

Hned pro začátek je nutno říct, že řádově 40% z celého objemu dat je značně vysoká hodnota. Abychom toto číslo ale lépe pochopili, je třeba se podívat na dílčí hodnoty IPD podrobněji.

Vzhledem k tomu, že celková hodnota IPD byla sumou sečtenou z různých výše popisovaných modelů (a u těch modelů, které byly založeny na jednotlivých výkonech, mohlo zařízení za různé druhy vykázaných výkonů obdržet body IPD za každý jeden z nich), byl výsledný součet IPD dosti rozmanitý. Konkrétně se pohyboval v uzavřeném intervalu od 1 do 186, kde hodnota 1 byla minimum pro zařazení do seznamu (hodnota 0 by představovala, že zařízení prošlo modely bez nálezu) a hodnota 186 pak byla maximem, kterého dosáhlo jedno zařízení.

Pokud se nyní zaměříme na rozdělení hodnot IPD na daném intervalu, zjistíme, že více jak 20 bodů IPD (což je zhruba prvních 10% daného intervalu) obdrželo „pouze“ 3 981 zařízení. Více jak polovina zdravotnických zařízení měla tedy 19 a méně bodů IPD. Když bychom pak pokračovali dál podobným způsobem, došli bychom k tomu, že hranici více jak 40 bodů IPD překročilo už jen 1 067 zdravotnických zařízení, což je něco málo přes 1/8 původního počtu zařízení.

Jak na číslech můžeme vidět, spolu s navyšující se hodnotou IPD počet zdravotnických zařízení podezřelých z neoprávněného vykazování úkonů markantně klesá. Názorně tento jev můžeme pak vidět na grafu na *obrázku 6 Počty hodnot IPD*. Ten zobrazuje, kolikrát byly jednotlivé hodnoty IPD zastoupeny ve výsledném souboru, který se předával zdravotní pojišťovně. Na grafu je jasně patrno, že spolu se snižující se hodnotou IPD, počet výskytů roste.



Obrázek 6 Počty hodnot IPD
Zdroj: vlastnoručně vytvořený graf z výsledných údajů

Pokud se pak zastavíme na hranici 100 bodů, zjistíme, že nad tuto hodnotu se dostalo pouhých 46 zařízení. Zde se tedy dostáváme k samotnému využití výstupů z datové analýzy v praxi.

Úkolem projektu pro pojišťovnu bylo najít zdravotnická zařízení, která s vysokou mírou pravděpodobnosti vykazují neoprávněné úkoly, protože provádět náhodné kontroly lékařů s pouze malou šancí nalézt pochybení je značně neefektivní a neekonomické.

I když cílem projektu nebylo výsledky datové analýzy dále zkoumat a nahlížet na ně z pohledu manažera, v bakalářské práci se na jejich možné praktické využití nyní podíváme.

Dopředu je nutné říct, že počty kontrol, které je pojišťovna schopna během jednoho roku z personálního hlediska provést, se pohybuje řádově v desítkách. Vezme-li tedy tuto skutečnost v potaz a podíváme-li se znovu na hranici 100 bodů, respektive na 46 nalezených zařízení s nejvyšší pravděpodobností možného podvodu, dostáváme se k okruhu těch lékařů, u kterých je pojišťovna schopna kontrolu uskutečnit. Zde je proto praktické využití výsledků datové analýzy zajištěno.

Dále se potom z manažerského pohledu nabízí použít výsledky analýz k preventivním účelům a obeslat zařízení, která se nacházela na intervalu IPD mezi 40 a 100 bodů, varovnými dopisy, neboť psychologický efekt podobného kroku často postačí k tomu, aby daná zařízení s neoprávněným vykazováním výkonů přestala.

2.5.2 Zhodnocení přínosů

Přínosy bakalářské práce by se daly rozdělit do dvou částí – do obecné a praktické roviny.

Co se týče té obecné, jak již bylo na samém začátku bakalářské práce řečeno, o využití pokročilých analýz dat v odvětví, jako je zdravotnictví, se v literatuře takřka nepíše. Proto by se jako jeden z hlavních a nejužitečnějších obecných přínosů práce dalo označit právě to, že tato bakalářská práce se zmíněnou problematikou (v okruhu vymezeném daným projektem) zabývá.

Praktické přínosy byly částečně zmíněny už v předchozí podkapitole. Jedná se především o přínos ekonomický, i když nemáme potřebná data k tomu, abychom ho byli schopni vyčíslit konkrétně. Jako další, neméně podstatný, bychom pak mohli uvést etický přínos, protože některá zdravotnická zařízení se obohacují zcela neoprávněně. Dělo a děje se tomu tak již dlouho a něco takového by se rozhodně tolerovat nemělo. Zvlášť když nyní máme vhodné prostředky, jako je právě data mining, ke zjišťování takových jednání.

Závěr

Hlavním přínosem bakalářské práce bylo praktické využití analýz dat spadajících do oblasti data miningu. Ty byly aplikovány pro projekt datové analýzy pro jednu ze zdravotních pojišťoven, která působí na českém trhu.

Cílem projektu bylo odhalit v datech poskytnutých klientem abnormality, které by indikovaly neoprávněně vykazovanou péči. Výsledkem prováděných analýz byla tabulka se sestupně seřazenými sumami IPD, tedy s vypovídající hodnotou od nejzávažnějších případů po ty, kdy je pravděpodobnost podvodu relativně nízká.

Výsledky získané z datové analýzy slouží jako podklad pro manažerské kroky zdravotní pojišťovny, která podle nich bude provádět cílené kontroly zdravotnických zařízení a lékařů. Tím se zároveň dostáváme k dalším přínosům bakalářské práce. Ty bychom mohly obecně rozdělit na přínos technický, ekonomický a společenský.

Jako technický přínos bakalářské práce bychom mohli označit výše popsaný, hlavní přínos. Co se ostatních dvou týče, ekonomickým přínosem je bezesporu to, kolik zdravotní pojišťovna díky tomuto projektu může ušetřit na výdajích za proplácení neoprávněně vykázaných úkonů.

Poslední přínos bakalářské práce leží v rovině společenské a spočívá v omezení podvodného chování zdravotnických zařízení. To se jistě bude v různých podobách objevovat i nadále, ale nyní díky data miningu bude snazší proti těmto praktikám bojovat.

Pokud bychom se pak zabývali otázkou, jak těchto metod dále využívat, odpovědí bychom nalezli hned několik. Jako první se nabízí aplikovat použité metody nejen takto na datech jedné pojišťovny, ale celoplošně i na údajích ostatních pojišťoven.

Další možností, jak by se daly metody data miningu dále rozšířit, je jejich aplikace například pro proplácení úhrad v nemocničních zařízeních. Ty v rámci popisovaného projektu nebyly do analýz zahrnuty, neboť jsou svou povahou specifické a je proto třeba k nim přistupovat zvlášť.

Jak je tedy jasně patrné, možností máme hodně. Teď už záleží pouze na manažerech samotných pojišťoven, zda i oni shledají výsledky datových analýz jako vhodný prostředek, jak mohou ušetřit.

Seznam použité literatury

Citace

[1] PRASANNA, D., H. KUO-WEI, J. SRIVASTAVA. Data mining for healthcare management. Arizona USA, 2011. Dostupné z: www.siam.org/meetings/sdm11/dmhm.pdf

[2] SEDLÁČKOVÁ, B. Data mining v knihovní a informační vědě. In: STEINEROVÁ, J. a J. ŠUSOL ed. *Využívanie informácií v informačnej spoločnosti*, 1. vyd. Bratislava: Filozofická fakulta Univerzity Komenského, 2006, s. 149-153. ISBN 80-85165-92-9. Dostupné z: www.academia.edu/2988265/O_autoroch

[3] Anonymní. Korelace. Dostupné z: http://meloun.upce.cz/docs/research/chemometrics/methodology/7metody.pdf

[4] VÍTEK. Členění data miningových úloh. Dostupné z: 2002http://datamining.xf.cz/view.php?cisloclanku=2002102801

[5] BERKA, P. *Dobývání znalostí z databází*. 1. vyd. Praha: Academia, 2003. ISBN 80-200-1062-9.

[6] GURINOVÁ, K. Analýza závislostí. Liberec. Dostupné z: http://multiedu.tul.cz/~katerina.gurinova/multiedu/Statistika_II/Analyza_zavislosti.pdf

[7] Anonymní. Diskriminační analýza. Dostupné z: http://cs.wikipedia.org/wiki/Diskrimina%C4%8Dn%C3%AD_anal%C3%BDza

[8] OUŘECKÁ, M. *Zpracování podkladů pro seminář předmětu PZDM v softwarovém prostředí Clementine – shluková analýza*. Pardubice, 2012. 51 s. Univerzita Pardubice, Fakulta ekonomicko-správní. Dostupné z: http://dspace.upce.cz/bitstream/10195/45618/2/OureckaM_Zpracovanipodkladu_TK_2012.pdf

[9] ACREA CR, spol. s.r.o. 2013. IBM SPSS Modeling family. Dostupné z: http://www.acrea.cz/cz/software/modeling-family

[10] SAS Institute Inc. Data Mining. Dostupné z: http://www.sas.com/offices/europe/czech/solutions/data_mining/index.html

[11] BÁRTÍK, F. Open-source nástroje pro data mining. 2011. Dostupné z: http://www.linuxexpres.cz/business/open-source-nastroje-pro-data-mining

[12] ORACLE. Oracle Advanced Analytics. Dostupné z: http://www.oracle.com/cz/products/database/options/advanced-analytics/index.html

Bibliografie

GROS, I. *Matematické modely pro manažerské rozhodování*. 1. vyd. Praha: Vydavatelství VŠCHT Praha, 2009. ISBN 978-80-7080-709-5.

NOVOTNÝ, O., J. POUR, D. SLÁNSKÝ. *Business Intelligence*. 1. vyd. Praha: Grada Publishing, 2005. ISBN 80-247-1094-3.

TUFFÉRY, S. *Data Mining and Statistics for Decision Making*. 1.vyd. Chichester: Wiley, 2011. ISBN 978-0-470-68829-8.

VENABLES, W. N., D. M. SMITH and R Core Team. *An Introduction to R* (Notes on R: A Programming Environment for Data Analysis and Graphics), Version 2.15.1 (2012-06-22). ISBN 3-900051-12-7.

Články dostupné z online zdrojů:

Elektronická databáze článků ProQuest (knihovna.tul.cz)

ŠIMÁNEK, R. Matematický software R: S ním je každá statistika hezčí. 2009. Dostupné z: http://www.linuxexpres.cz/software/matematicky-software-r-s-nim-je-kazda-statistika-hezci